UNIVERSITA' DI PISA

FACOLTA' DI MEDICINA E CHIRURGIA

Scuola di Specializzazione
in
Patologia Clinica

L'evoluzione del sequenziamento del DNA: dal metodo di Sanger alle tecnologie NGS

Relatore
Prof. Aldo Paolicchi

Candidato
Annamaria Marcocci

Anno Accademico 2013/2014

SOMMARIO

RIASSUNTO *pag 4*

INTRODUZIONE

- Il Genoma Umano *pag 7*
- Il Progetto Genoma Umano *pag 11*
    - L'era post-genomica *pag 13*

IL SEQUENZIAMENTO

- Concetti Generali *pag 14*
- Storia del sequenziamento *pag 14*
- Il sequenziamento di prima generazione *pag 17*
    - Metodo di Maxam-Gilbert *pag 17*
    - Metodo di Sanger *pag 19*
- NGS: Metodiche di 2° Generazione *pag 21*
    - Tecnologia 454 *pag 24*
    - Illumina/Solexa *pag 26*
    - Applied Biosystems: SOLiD *pag 29*
    - Ion Torrent *pag 32*
- NGS: Metodiche di 3° Generazione *pag 33*
    - PacBio RS *pag 35*
    - Heliscope Sequencer *pag 37*
- NGS: Metodiche di 4°Generazione *pag 38*
    - Tecnologia a nanopori *pag 38*
- Analisi dei dati *pag 40*

AMBITI DI APPLICAZIONE DELLE TECNICHE NGS

- Analisi genomica *pag 44*
- Risequenziamento *pag 45*
- Analisi del trascrittoma *pag 46*
- Metagenomica *pag 48*
- Farmacogenetica e farmacogenomica *pag 49*

CONCLUSIONI *pag 51*

BIBLIOGRAFIA *pag 54*

RIASSUNTO

INTRODUZIONE: Un gene è una parte di genoma umano che contiene tutte le informazioni necessarie alla produzione, mediante trascrizione e traduzione, di una proteina funzionale; la definizione più completa e precisa di gene è stata formulata da Mark Gerstein nel 2007 "*un gene è l'unione di sequenze genomiche che codificano per un set coerente di prodotti funzionali potenzialmente sovrapponibili",* questa definizione tiene conto di tutte le accezioni di gene.

In seguito al completamento del Progetto Genoma Umano, i cui risultati furono pubblicati nel 2003, sono stati identificati circa 20.000–25.000 geni. Alcuni ricercatori hanno fornito una stima precisa del numeri di geni presenti secondo il modello di Craig Venter (nel 2007), asserendo che i geni sarebbero 23.224, mentre secondo Jim Kent (2007) sarebbero 20.433 codificanti e 5.871 non codificanti.

Un aspetto che è stato evidenziato dal progetto genoma umano è l'osservazione che la parte del genoma che viene tradotta in proteina rappresenta solo il 1,5% di tutto il materiale genetico, un altro 1,5% è rappresentato da sequenze regolatrici conosciute ed introni, ma il 97% di tutto il DNA rimane a funzione ignota.

Il sequenziamento del DNA acquisisce quindi un ruolo fondamentale con conseguente necessità di tecniche che riducono costi e tempi di lavoro, ottenendo risultati sempre più attendibili.

SEQUENZIAMENTO: Con il termine sequenziamento si intende la decodificazione dell'esatta sequenza dei nucleotidi di un acido nucleico (DNA o RNA).

I primi modelli di sequenziamento furono elaborati da Maxam e Gilbert e da Sanger nel 1977, mentre dal 2005 sono state messe in commercio le prime piattaforme di Next Generation Sequencing (NGS).

I modelli di sequenziamento di Maxam e Gilbert e il modello di Sanger sono detti tecnologie di prima generazione, mentre le tecnologie NGS sono state classificate in tecnologie di seconda, terza e quarta generazione.

APPLICAZIONI IN DIAGNOSTICA E RICERCA: L'avvento delle tecnologie NGS ha notevolmente accelerato la crescita di vari settori di ricerca genomica, consentendo di effettuare esperimenti che in precedenza presentavano notevoli ostacoli soprattutto da un punto di vista economico.

In particolare sono stati potenziati gli studi di analisi genomica, sequenziamento de novo, analisi del trascrittoma, sostituzione della tecnica dei microarray per la mappatura della cromatina; inoltre ha permesso l'evolversi di settori di ricerca quali la metagenomica, la farmacogenomica e la farmacogenetica.

CONCLUSIONI: Dalle prime tecniche di sequenziamento sono stati fatti notevoli passi avanti e ancora nuove tecnologie sono in fase di sviluppo (Nanopore Technologies) con l'obiettivo di ridurre i tempi e i costi ed aumentare la conoscenza del genoma.

I principali obiettivi verso cui le ricerche si sono indirizzate sono un ulteriore

miglioramento delle tecnologie già presenti e la possibilità che piattaforme NGS possano essere utilizzate in maniera routinaria nella diagnostica clinica.

INTRODUZIONE

IL GENOMA UMANO

Con il termine genoma umano si descrive l'insieme di tutto il DNA che costituisce il patrimonio genetico della specie "Homo Sapiens"; questa enorme quantità di DNA è suddivisa in 23 coppie di cromosomi omologhi e si ripartisce in *geni*, *sequenze regolatrici* e un insieme di elementi non codificanti, a funzione ancora non definita, tra cui troviamo *elementi ripetuti*, *trasposoni*, *pseudogeni.*

Un gene è una parte di genoma umano che contiene tutte le informazioni necessarie alla produzione, mediante trascrizione e traduzione, di una proteina funzionale; la definizione più completa e precisa di gene è stata formulata da Mark Gerstein nel 2007 "*un gene è l'unione di sequenze genomiche che codificano per un set coerente di prodotti funzionali potenzialmente sovrapponibili*" [1], questa definizione tiene conto di tutte le accezioni di gene.

Un gene procariote è costituito solo da sequenze di DNA che vengono tradotte in proteine (sequenze codificanti) mentre nel gene eucariotico troviamo un insieme di sequenze codificanti, definite esoni, e sequenze non codificanti, definite introni, che possono avere una funzione regolatoria, a ciò si aggiunge un promotore, una zona di DNA non tradotta deputata al controllo dell'espressione genica.

Durante la trascrizione esoni e gli introni sono trascritti da DNA ad RNA formando un pre-mRNA così definito poiché immaturo; successivamente ad esso vengono rimossi gli introni mediante un processo definito splicing (in molti casi si ha uno splicing alternativo, che permette alla cellula di sintetizzare più proteine a partire da un unico trascritto primario); viene poi aggiunto un

cappuccio guanosinico, che ne impedisce la degradazione e una coda poliadenilica, anch'essa coinvolta nella protezione del trascritto; cosi si origina l'RNA messaggero (o mRNA), il quale dirige la sintesi delle proteine.

I geni dirigono lo sviluppo fisico e comportamentale di un essere vivente, in quanto la maggior parte di essi codifica per proteine, le macromolecole maggiormente coinvolte nei processi biochimici e metabolici della cellula.

Ogni singolo cambiamento nella sequenza del DNA costituisce una mutazione e può causare una conseguente alterazione nella sequenza di amminoacidi di una proteina o nella regolazione della sua espressione (che, in conseguenza, potrebbe anche avere conseguenze patologiche). È stato calcolato che le alterazioni dei nostri geni sono responsabili di circa 5000 malattie ereditarie (per esempio vari tipi di anemia). Altre mutazioni, anziché evidenziarsi in maniera diretta come malattia, possono causare una predisposizione ad esse.

In base ad un loro aspetto funzionale, si definiscono **geni strutturali** quei geni che codificano per una proteina la cui funzione principale è la costituzione di una struttura fisica all'interno di una cellula. Essi determinano l'ordinata successione di amminoacidi nella catena polipeptidica sulla base delle proprie sequenze di basi mentre si definiscono **geni regolatori** quei geni che contengono le informazioni relative a molecole che regolano l'espressione di altri geni (geni strutturali); come ad esempio i geni omeotici

In seguito al completamento del **Progetto Genoma Umano** [2], i cui risultati furono pubblicati nel 2003, sono stati annoverati circa 20.000–25.000 geni. Alcuni ricercatori hanno fornito una stima dei numeri di geni presenti secondo

il modello di Craig Venter (nel 2007), asserendo che i geni sarebbero 23.224, mentre secondo Jim Kent (2007) sarebbero 20.433 codificanti e 5.871 non codificanti.[3,4]

Le sequenze regolatrici, invece, sono solitamente brevi sequenze che si trovano solitamente in prossimità o all'interno dei geni ed hanno la funzione base di controllare l'espressione del gene.

Un aspetto che è stato evidenziato dal progetto genoma umano è l'osservazione che la parte del genoma che viene tradotta in proteina rappresenta solo l'1,5% di tutto il materiale genetico, un altro 1,5% è rappresentato da sequenze regolatrici conosciute ed introni, ma il 97% di tutto il DNA rimane a funzione ignota.[2]

Alcune zone di DNA per le loro caratteristiche strutturali e funzionali sono state classificate strutture a funzione ignota e si suddividono in elementi ripetuti, trasposoni e pseudogeni.

Gli elementi ripetuti sono brevi sequenze di basi ripetute all'interno del genoma si dividono in ripetizioni in tandem e ripetizioni intersperse.

Le ripetizioni in tandem sono brevi sequenze di basi ripetute una di seguito all'altra, si dividono in DNA satellite, DNA minisatellite e DNA microsatellite, sono strutture ad elevato polimorfismo che si rivelano fondamentali negli studi di parentela e nei test di confronto di DNA.

Le ripetizioni intersperse invece sono separate lungo il genoma e in base alla lunghezza si suddividono in SINE *''short interspersed nuclear element''* e LINE, *''long interspersed nuclear element''*.

Le ripetizioni intersperse rappresentano anche un tipo di elemento trasponibile, ovvero una sequenza di DNA che può muoversi all'interno del genoma.

Gli elementi trasponibili umani sono solitamente retrotrasposoni, cioè sequenze che per essere trasposte devono necessariamente essere convertite da una trascrittasi inversa in DNA, più rari sono i trasposoni a DNA.

Gli pseudogeni sono copie non funzionali di DNA genomico, strutturalmente sono simili a un gene convenzionale essendo spesso costituiti da esoni, introni e sequenze regolatori ma l'analisi della sequenza evidenzia la presenza di codoni di stop all'interno degli esoni che portano inevitabilmente alla formazione di proteine non funzionali.

Possono essere sia trascritti che tradotti ma non possono produrre una proteina funzionale oppure possono essere frammenti tronchi di un gene e quindi non essere in grado di permettere trascrizione e traduzione.

Tuttavia vi è ancora una grande quantità di sequenze di DNA che non può essere compreso in nessuno di questi elementi, parte di questo DNA viene trascritto ma non se ne conosce la funzione, parte invece non viene trascritto e non se ne conosce ancora la funzione.

IL PROGETTO GENOMA UMANO

Il **"Progetto Genoma Umano"** (Human Genome Project HGP) è un progetto realizzato dall'U.S. Department of Energy (DOE) e dal National Insitute of

Health con la partecipazione di strutture di Tecnologia e Ricerca di tutto il mondo; il progetto è durato 13 anni dal 1990 al 2003 ed aveva come obiettivo la determinazione della sequenza dell'intero genoma umano.[2]

Gli obiettivi principali di questo progetto erano di identificare tutti i geni presenti nel genoma umano, definire il corretto ordine di tutti i 3 miliardi di basi presenti nel genoma e creare un database in grado di raccogliere queste informazioni da mettere a disposizione di enti di ricerca pubblici e privati per studi successivi.

Inizialmente si trattava di un progetto pubblico guidato dalle due istituzioni americane ma successivamente il ricercatore Grag Venter lasciò il National Institute of Health per creare un ente di ricerca privato la Celera Genomics portando avanti parallelamente il progetto di caratterizzazione del genoma umano.

La metodica di sequenziamento utilizzata era il metodo enzimatico di Sanger, la spinta iniziale al progetto fu data proprio dall'introduzione dei primi sequenziatori automatici da parte della ditta Applied Biosystems nel 1986 che permettevano di codificare circa 400000 basi al giorno riducendo notevolmente i tempi necessari all'ottenimento di una sequenza.

Proprio grazie alla riduzione dei tempi necessari alla sequenza che la ditta Celera Genomics ha potuto elaborare un approccio al sequenziamento definito "shot-gun".

L'approccio di sequenziamento "shotgun" si basa sul clonaggio di frammenti multipli di piccole dimensioni per poi ricostruire la sequenza di contigui

definitiva

I primi risultati furono pubblicati contemporaneamente nel 2001 sia dall'ente pubblico che dal privato e riportavano minime differenze. [5]

La pubblicazione della sequenza di un intero genoma umano aploide avvenne nel 2003, successivamente nel 2011 è stato codificato il primo genoma diploide appartenente a James Watson.

I risultati del progetto genoma umano non furono cosi eclatanti come si prevedeva, il primo dato che venne evidenziato fu che il genoma umano è costituito da circa 20500 geni e non 100000 mila come si prevedeva prima dell'inizio del progetto, mettendo quindi in luce le esigue differenze in numero di geni tra un essere umano e alcuni organismi modello come il nematode Caenorhabditis elegans (19000geni identificati) o l'artropode Drosophila Melanogaster (13000).

Parallelamente a questo è stato evidenziato come solo l'1,5% del nostro DNA è tradotto in proteine mentre per il resto si tratta di sequenze regolatrici che influiscono in maniera diversa sull'espressione proteica e sulle funzioni cellulari.

L'ERA POST-GENOMICA

I risultati ottenuti dal sequenziamento del genoma umano hanno evidenziato un concetto fondamentale: non si può pensare che la sola sequenza nucleotidica possa essere sufficiente a spiegare le funzioni biologiche dell'organismo e le

differenze tra organismi di complessità diversa, ciò che genera la diversità tra gli organismi sono i meccanismi post-trascrizionali e post-traduzionali.

La presa di coscienza di questa nuova visione del patrimonio genetico idealmente chiude quindi quella che venne definita l'era genomica per lasciare spazio all'era post- genomica.

Gli studi di post-genomica sono quindi più indirizzati allo studio del trascrittoma, ovvero la porzione di DNA che viene realmente trascritta e diventa RNA funzionale e del proteoma cioè la parte di RNA che viene tradotta in proteine; inoltre diventano sempre più importanti i ruoli degli SNP (single nucleotide polimorphism) e delle VNTR (variable number tandem repeat).

Quando fu lanciato il progetto per sequenziare il genoma umano, si conoscevano meno di 100 geni legati alle malattie; attualmente sono stati definiti più di 2850 geni relativi a malattie di tipo mendeliano.

Avere la possibilità di leggere un intero genoma diventa sempre più fondamentale anche per la pratica clinica: per identificare le cause genetiche di malattie rare particolarmente difficili da diagnosticare, molti medici iniziano a scegliere la strada del sequenziamento genomico per la diagnosi molecolare.

Il SEQUENZIAMENTO

CONCETTI GENERALI

Con il termine sequenziamento si intende la decodificazione dell'esatta sequenza di nucleotidi di un acido nucleico (DNA o RNA).

Un filamento di DNA o RNA contiene qualche miliardo di nucleotidi, ma i dispositivi di analisi hanno una capacità limitata di lettura, permettendo di sequenziare solo dei frammenti, di lunghezza variabile, denominati "reads"; le "reads" devono poi essere allineate e assemblate per formare la sequenza di nostro interesse.

STORIA DEL SEQUENZIAMENTO

Nel 1953 Watson e Crick caratterizzarono il DNA nella loro struttura, fornendo il punto di partenza per la comprensione del codice genetico e della trasmissione delle informazioni, da questa scoperta inizio a manifestarsi la necessità di conoscere e decifrare il DNA.

Il primo esempio di sequenziamento è stato nel 1973[6] con l'elaborazione da parte di Maxam e Gilbert di un breve frammento di circa 24bp e successivamente nel 1977[7] pubblicarono la loro metodica di sequenziamento.

Sempre 1977 Sanger pubblica la sua metodica di sequenziamento enzimatico, questa tecnologia fu successivamente automatizzata e commercializzata rimanendo per molti anni il "gold standard" per il sequenziamento del DNA sia in campo diagnostico che di ricerca.[8]

La metodica di Sanger fornì la possibilità di sequenziare frammenti più lunghi di DNA, ciò portò nel 1982 alla creazione del progetto GenBank un database creato da una sezione del National Institute of Health, il National Center for Biotechnology Information (NCBI), allo scopo di raccogliere e mettere a

disposizione di tutti le sequenze elaborate.[9]

Nel 1983 fu introdotta la metodica della PCR[10] e nel 1986 la ditta Applied Biosystem introdusse nel mercato il primo sequenziatore automatico[11], queste due innovazioni dettero una spinta fondamentale all'avvio del progetto Genoma Umano iniziato nel 1990 e terminato nel 2003.

Nel 1998 fu introdotto nel mercato il primo sequenziatore automatico a capillare (ABI 310)[11] e nello stesso anno fu sequenziato il primo genoma completo del Caenorhabditis elegans[12] e nel 2003 il primo genoma umano aploide.

Nel 2005 si apre l'era della Next Generation Sequencing con l'introduzione nel mercato della piattaforma GS-20 System per merito della 454 Life Sciences, [13] nel 2006 sarà invece la Solexa ad introdurre la piattaforma Genome Analyzer[14] e l'anno successivo nel 2007 sarà la volta del sistema Solid dell'Applied Biosystems.[15]

Nel 2009 la ditta Helicos immette sul mercato la piattaforma Helicos Genetic Analyser Sistem che elabora un nuovo metodo di sequenziamento in "Single Molecule" basato su un sistema già ideato nel 2003 da Brasvlasky [16,17], seguito nel 2011 dalla Pacific Bioscences con la piattaforma PacBIO RS Sistem, nello stesso anno la Ion Torrent immette nel mercato la piattaforma PGM.

Nel 2012 viene mostrata per la prima volta la Nanopore Technologies per il sequenziamento diretto del DNA[18], la piattaforma non è ancora in commercio.

Appare chiaro come negli ultimi 10 anni le tecnologie di sequenziamento si siano notevolmente evolute, perseguendo l'obiettivo di abbassare i costi, ridurre i tempi di lavoro e gli step di processazione cosi da minimizzare la possibilità di introdurre errori.

Il sistema di Sanger è un sistema di sequenziamento processivo che consente il sequenziamento di un singolo campione di DNA, caratterizzato da alti costi di esecuzione, tempi lunghi di processazione e uno scarso throughput, invece le tecniche di Next Generation Sequencing sono caratterizzate dalla possibilità di eseguire un sequenziamento parallelo e massivo, permettendo di poter analizzare in una seduta fino a 1Gb.

Le tecniche NGS hanno notevolmente ridotto i costi e i tempi necessari all'elaborazione di una sequenza, rendendo possibili tanti studi che prevedono l'analisi di numerose sequenze di DNA.

Uno svantaggio di queste tecniche è la ridotta lunghezza delle "reads" che rendono più complesso il processo di allineamento e assemblaggio, generando in questa fase alcuni errori di sequenza.

A seguito della rapida evoluzione nelle tecnologie NGS, si è tentato di fare una classificazione dalle tecnologie di tipo NGS differenziandole in tecniche di seconda generazione (2G) di terza generazione (3G) e le più recenti tecnologie di quarta generazione (4G).[19]

Si tratta di una classificazione non stringente che raggruppa le tecnologie di sequenziamento sulla base di caratteristiche peculiari nello svolgimento della metodica e sull'ordine cronologico di elaborazione

IL SEQUENZIAMENTO DI PRIMA GENERAZIONE

Il sequenziamento di prima generazione comprende il metodo di Maxam-Gilbert e di Sanger che sono stati i primi due metodi ad essere stati approntati ed utilizzati.

Tra i due, il metodo ideato da Sanger (o metodo enzimatico), a seguito dell'avvento dei sequenziatori a capillari che hanno automatizzato la procedura di interpretazione dei dati ed hanno abbattuto i tempi necessari all'ottenimento della sequenza di DNA, rimane la metodica "gold standard" in molti laboratori sia di analisi che di ricerca.

Metodo di Maxam-Gilbert

Il metodo Maxam-Gilbert fu per la prima volta descritto nel 1977 da parte dei ricercatori Allan Maxam e Walter Gilbert e si basa su alterazioni chimiche del DNA e sul conseguente taglio in posizioni specifiche

La metodica prevede che il DNA da sequenziare sia purificato e marcato radioattivamente ad un'estremità (generalmente usando 32P). Il campione di DNA da sequenziare viene denaturato in presenza di DMSO e viene diviso in quattro aliquote uguali, ciascuna delle quali viene trattata con dei reagenti chimici che ne causano la metilazione o la rottura in corrispondenza di basi specifiche.

Utilizzando i reagenti a basse concentrazioni si può fare in modo che i tagli non avvengano su tutte le basi ma, ipoteticamente, un taglio per ogni molecola di DNA cosicché viene generata una serie di frammenti marcati (dalla fine della molecola al primo sito di taglio della stessa) di dimensione specifica che vengono separati in base alla lunghezza attraverso elettroforesi su gel.

Il gel viene posto a contatto con una pellicola radiografica sulla quale lascia impressa la disposizione delle bande dei frammenti generati tramite i quali è possibile determinare l'ordine dei nucleotidi e quindi la sequenza di partenza. [6]

Il metodo sviluppato da Maxam e Walter Gilbert viene comunemente descritto come metodo chimico per differenziarlo dal metodo enzimatico di Sanger. Questo metodo originò da studi sulle interazioni DNA-proteine (footprinting), sulla struttura degli acidi nucleici e su modificazioni epigenetiche, e in questi campi la metodica ha tuttora applicazioni importanti. Sebbene i due pubblicarono questa tecnica due anni dopo la pubblicazione di Sanger, il loro metodo divenne immediatamente popolare e preferito, poiché il DNA purificato poteva essere utilizzato direttamente, senza passare per un intermedio a singolo filamento, come invece richiesto dal metodo dei terminatori di catena. Comunque, con il successivo miglioramento del metodo Sanger, il Maxam-Gilbert venne progressivamente accantonato a causa della complessità tecnica e dell'uso estensivo di sostanze tossiche, oltre al fatto che si è dimostrato piuttosto difficile poter sviluppare un kit da laboratorio pronto all'uso.

Metodo di Sanger

Il metodo Sanger è un metodo cosiddetto enzimatico, poiché richiede l'utilizzo di un enzima; questa tecnica si basa sull'utilizzo di nucleotidi modificati (dideossitrifosfato, ddNTPs) per interrompere la reazione di aggiunta di nucleotidi in posizioni specifiche.

I nucleotidi dideossitrifosfato sono molecole artificiali corrispondenti ai nucleotidi naturali, ma si differenziano per l'assenza del gruppo idrossilico sul carbonio 2' e 3' della molecola. I dideossinucleotidi, a causa della loro conformazione, impediscono che un altro nucleotide si leghi ad essi, in quanto non si possono formare legami fosfodiesterici.

Il protocollo classico richiede un template di DNA a singolo filamento, un primer per iniziare la reazione di polimerizzazione, una DNA polimerasi, deossinucleotidi e dideossinucleotidi per terminare la reazione di polimerizzazione.

I nucleotidi modificati (ddNTPs) sono marcati con un fluroforo in modo da poter visualizzare le bande dei frammenti di DNA neosintetizzato dopo aver effettuato l'elettroforesi.

Nella metodica iniziale il materiale genetico da sequenziare veniva diviso in quattro reazioni separate, ognuna delle quali contenente la DNA polimerasi e tutti e 4 i deossiribonucleotidi (dATP, dCTP, dGTP, dTTP). Ad ognuna di queste reazioni veniva poi aggiunto solo uno dei quattro dideossinucleotidi

(ddATP, ddCTP, ddGTP, ddTTP) in quantità stechiometricamente inferiore per permettere una elongazione del filamento sufficiente per l'analisi. L'incorporazione di un dideossinucleotide lungo il filamento di DNA in estensione ne causa la terminazione prima del raggiungimento della fine della sequenza di DNA stampo; questo dà origine ad una serie di frammenti di DNA di lunghezza diversa interrotti in corrispondenza dell'incorporazione del dideossinucleotide, che avviene casualmente quando esso è aggiunto dalla polimerasi in luogo di un nucleotide deossi

I frammenti generati da queste reazioni venivano poi fatti correre su gel di poliacrilamide-urea in grado di separare frammenti che differiscono anche di una sola base. Ognuna delle 4 reazioni è corsa su pozzetti vicini, dopodiché le bande sono visualizzate su lastra autoradiografica o sotto luce UV, e la sequenza viene letta direttamente sulla lastra o sul gel, a seconda del tipo di marcatura dei nucleotidi dideossi[7]

Basandosi su questa procedura, la metodica è stata affinata per facilitare la reazione, e con l'avvento dell'automatismo, la reazione di sequenziamento è diventata molto più veloce.

Attualmente è possibile effettuare, anziché quattro reazioni distinte per ogni nucleotide modificato, una sola reazione utilizzando i 4 ddNTPs marcati fluorescentemente in modo diverso tra loro ed utilizzando lettori ottici appropriati. Al posto del gel di elettroforesi in poliacrilamminde-urea viene utilizzato un capillare contenente un polimero (POP) all'interno del quale i frammenti vengono fatti correre secondo le loro dimensioni; all'estremità

opposta del capillare è presente un laser che eccitta il fluoroforo che emette una luce diversa a seconda del ddNTP incorporato, il sistema di rilevazione interpreta il segnale e elabora la sequenza.

La messa in commercio di kit per il sequenziamento pronti all'uso e di sequenziatori a capillare automatici (attualmente esistono sequenziatori a 24 capillari) ha reso per molto tempo il metodo di Sanger il "gold Standard" per il sequenziamento ed è ancora oggi utilizzato in molti laboratori clinici e di ricerca.

NGS:METODICHE DI 2° GENERAZIONE

I metodi di 2° generazione (2G) sono stati i primi metodi di sequenziamento parallelo e massivo introdotti nel mercato, questi metodi hanno permesso di abbattere notevolmente i tempi di lavoro e i costi, attualmente infatti il sequenziamento di un intero genoma richiede una spesa intorno ai 2000 dollari.

Esistono diversi metodi di sequenziamento 2G che si basano su processi biochimici differenti, ma esistono tuttavia delle linee guida comuni sulla procedura.*(Figura 1)*

Il primo passo consiste infatti nella preparazione di una libreria di DNA che solitamente si ottiene per frammentazione del DNA di interesse; ai frammenti cosi ottenuti vengono aggiunti degli "adaptors", ovvero delle sequenze specifiche di poche basi, necessarie ad ancorare e immobilizzare i frammenti sul supporto dove avverrà la successiva amplificazione e il sequenziamento.

La fase successiva è l'amplificazione necessaria per aumentare il numero di copie di ogni singolo frammento di DNA (generazione di cluster), la PCR può avvenire o in emulsione o in soluzione; le copie clonali di ogni frammento prendono il nome di clusters.

La generazione dei cluster è un passaggio cruciale in quanto, poiché la velocità di crescita e sequenziamento dei frammenti deve essere la stessa, è necessario limitare la lunghezza delle singole "reads".

Terminata la fase di amplificazione, segue il sequenziamento vero e proprio e l'immagazzinamento e l'analisi dei dati.

Adesso vediamo una panoramica delle principali tecnologie in ordine cronologico di elaborazione

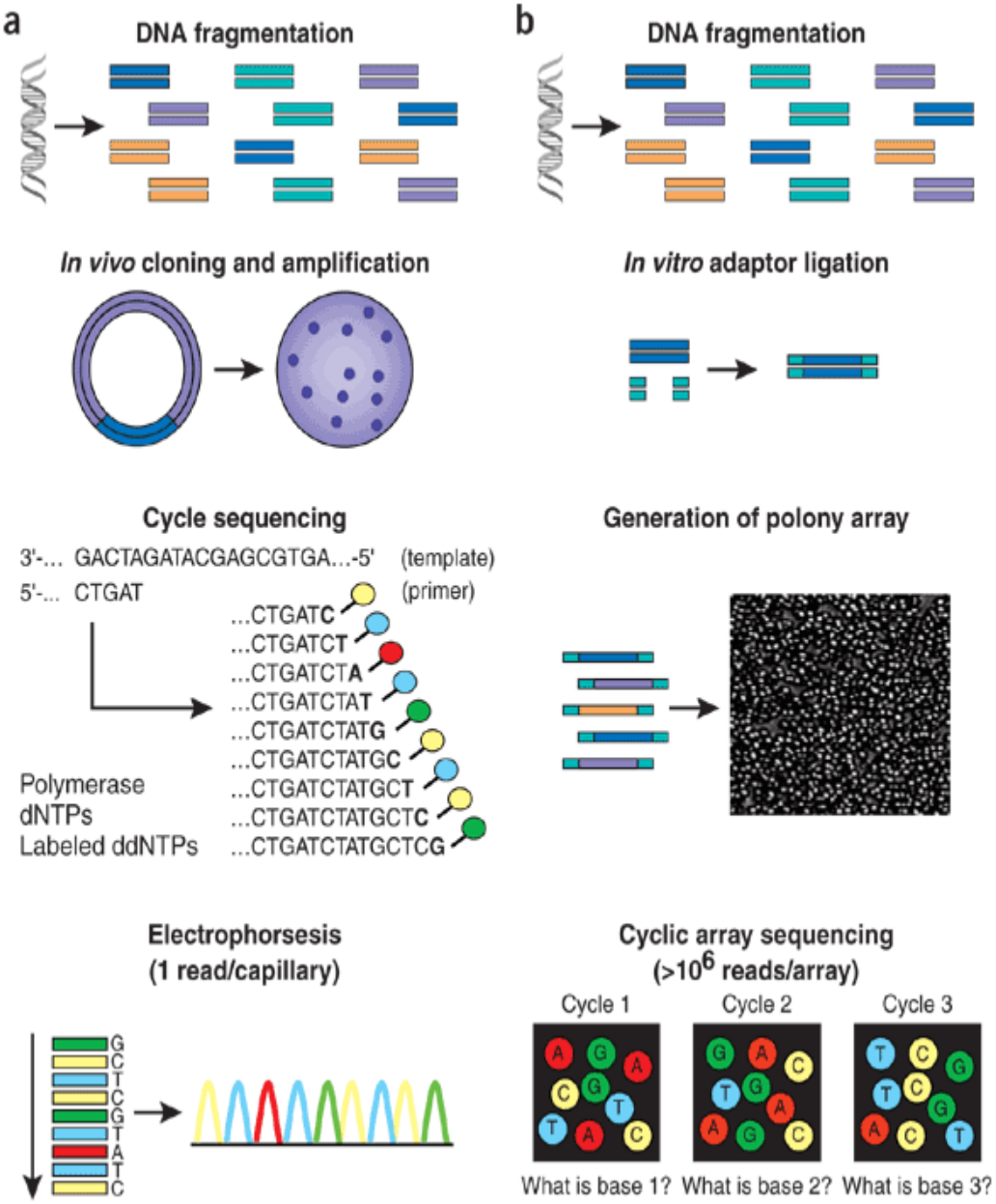

Figura 1:Confronto tra il metodo di Sanger(a) e i metodi NGS(b) (*Next Generation Sequencing.Shandure,Janleee Nature Biotechnology)*

Tecnologia 454

La tecnologia 454 nasce dalla convergenza di due metodiche: il pirosequenziamento descritto nel 1993 da Nyren [20]e la PCR in emulsione progettata da Tawfik e Griffiths nel 1998 [21]; nel 2000 Jonathan Rothberg fonda la 454 Life Sciences e nel 2005 immette in commercio la prima piattaforma per il sequenziamento la GS20, nel 2007 la società è acquisita da Roche Applied Science e verrà messa in commercio una successiva piattaforma la GS FLX.

La metodica per queste due piattaforme è molto simile, la prima fase prevede la frammentazione del DNA stampo mediante sonicazione o nebulizzazione, questa frammentazione porta alla formazione di frammenti di DNA a doppio filamento lunghi alcune centinaia di basi a cui vengono legati oligonucleotidi adattatori.

Successivamente le singole molecole di DNA sono denaturate e ibridate a singole biglie rivestite da sequenze complementari a quelle degli adattatori, quindi le biglie ricoperte di frammenti di DNA sono catturate in goccioline di emulsione acqua-olio in cui avviene la reazione di amplificazione clonale, dopodiché sono depositate nei singoli pozzetti di una piastra "picotiter" e combinate con i componenti necessari per la reazione di sequenziamento.

La fase di sequenziamento prevede l'utilizzo di un primer di innesco e la reazione dei frammenti con soluzioni contenenti singoli dNTP, se viene incorporato un nucleotide la reazione chimica prevede il rilascio di un

pirofosfato che viene convertito in ATP dall'ATP solforasi; l'ATP cosi formato agisce da carburante per convertire la luciferina in ossiluciferina che è in grado di generare una luce visibile proporzionalmente all'ATP presente e quindi di conseguenze al numero di basi incorporate.

L'enzima apirasi degrada i nucelotidi non incorporati e l'ATP, per permettere di testare con i nucleotidi successivi.

Il segnale luminoso derivante da ciascun pozzetto viene acquisito, analizzato per ridurre i rumori di fondo e in seguito elaborato per produrre una sequenza lineare.*(Figura 2)*

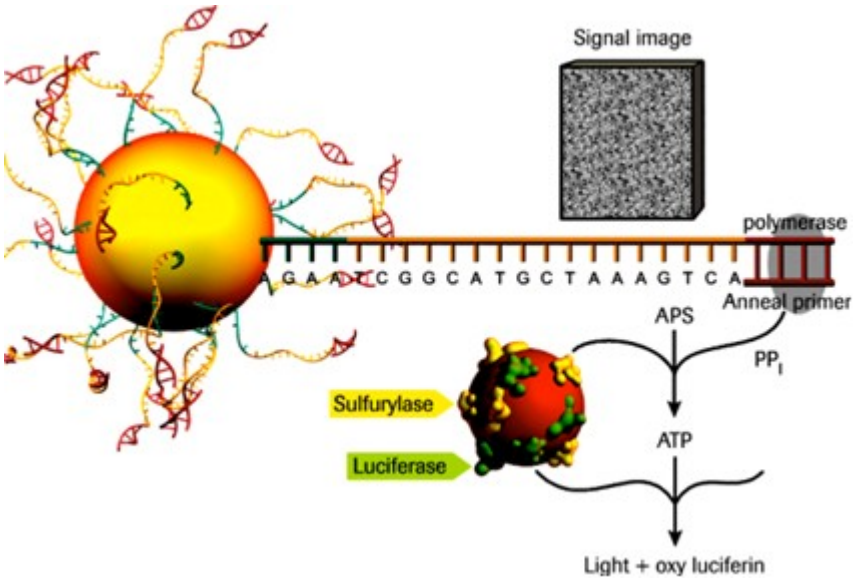

Figura x:Flusso di lavoro tecnologia 454*(BiotechLand2007-2014)*

In termini numerici una singola seduta analitica può sequenziare circa 500 milioni di bp, un punto di forza di questa tecnologia è la lunghezza delle "reads" maggiore rispetto ad altri sistemi rendendo estremamente agevole il

sequenziamento *de novo*. [22]

Nonostante l'abbattimento dei costi, rispetto al sequenziamento di Sanger, questa metodica rimane la più costosa di quelle presenti in commercio e mantiene, nonostante lo sviluppo e il perfezionamento, un margine di errore che oscilla intorno all'1,07% con maggiore frequenza in particolari posizioni genomiche.[23]

Il problema maggiore di questa metodica si rileva negli omopolimeri, infatti quando avviene l'incorporazione di più nucleotidi identici consecutivi il segnale dovrebbe essere proporzionale alla quantità di nucleotidi aggiunti (1 nucleotidi=1 segnale, 6 nucleotidi=6x1 segnale) in realtà il segnale rileva una variazione che però non viene sempre percepita nelle quantità esatte.

Illumina/Solexa

La piattaforma *Genome Analyzer* è stata introdotta sul mercato nel 2006 dall'azienda Solexa, successivamente acquisita da Illumina, attualmente la ditta Illumina produce anche la piattaforma *HiSeq* e *MiSeq*.

Come nella tecnologia 454 il DNA campione viene frammentato e processato al fine di ottenere estremità tronche fosforilate al 5', mentre l'attività enzimatica del frammento di Klenow aggiunge una singola base di adenina al 3'del DNA stampo.

Questa aggiunta ha il compito di facilitare l'attacco degli oligonucleotidi adattatori, i quali sono a loro volta complementari con oligonucleotidi ancorati

alla base della cella a flusso.

L'amplificazione clonale dei frammenti avviene, a differenza della PCR in emulsione, secondo un procedimento di amplificazione a ponte che si basa sulla cattura dei filamenti di DNA ripiegati ad arco che si ibridizzano ad un oligonucleotide adiacente.

Cicli multipli di amplificazione convertono la singola molecola stampo in un cluster di circa 1000 ampliconi, in una singola cella di flusso possono essere generati circa $50x10^6$ cluster separati.

Per il sequenziamento, ogni singolo amplicone è denaturato e successivamente vengono rimossi con un lavaggio i filamenti antisenso, mentre la sequenza sul filamento senso (forward) avviene mediante l'ibridazione di un primer complementare alla sequenza dell'oligonucleotide adattatore a cui viene aggiunta una miscela di nucleotidi marcati con fluorofori differenti.

Ogni ciclo di sequenza prevede che gli ampliconi vengono testati con una miscela di DNA polimerasi e i dNTP a cui è aggiunto un fluoroforo e un terminatore reversibile, il terminatore per caratteristica impedisce che venga aggiunto un nucleotide successivo, la sua reversibilità fa si che possa essere rimosso per permettere l'aggiunta di un nucleotide nel ciclo di sequenziamento successivo.

Dopo ogni incorporazione di dNTP, il laser eccita il fluoroforo e permette di identificare la base incorporata, dopodiché un lavaggio rimuove fluoroforo e terminatore permettendo l'aggiunta di una base successiva.[14]*(Figura 3)*

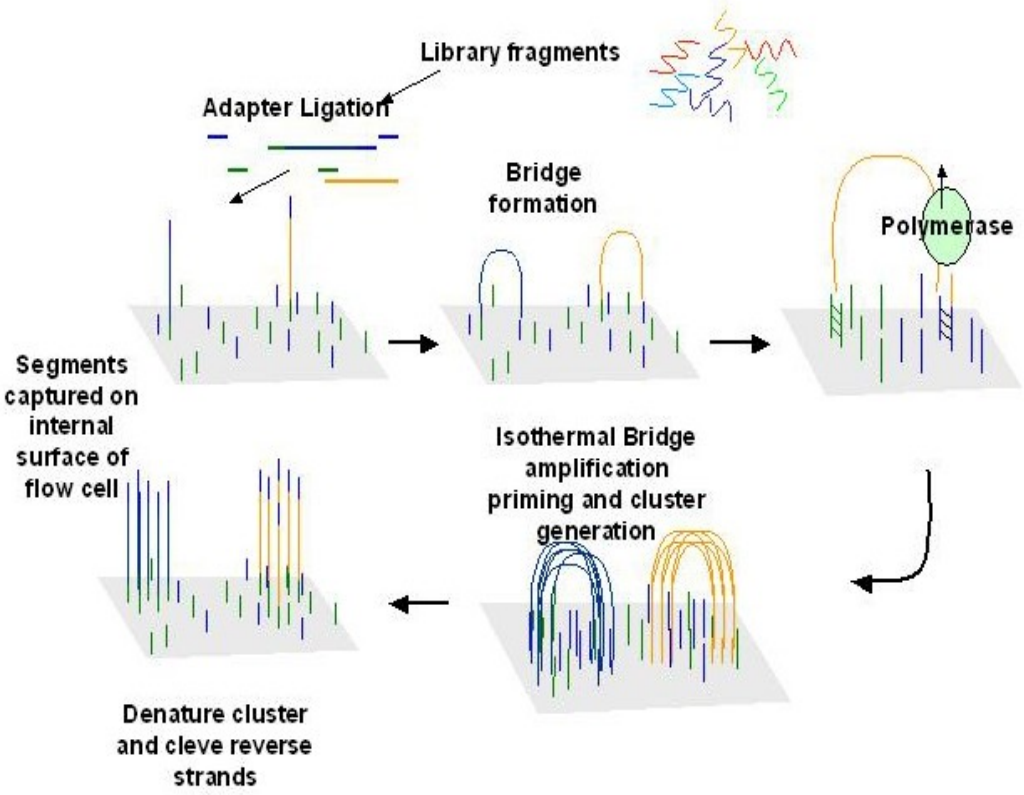

Figura 3:Flusso di lavoro tecnologia Illumina (*Ken Laing IPC, Centre for Infection and Immunity Division of Clinical Sciences, St George's University of London*

Attualmente i sequenziatori Genoma Analyzer possono leggere fino a 6 miliardi di "reads" in un ciclo di sequenziamento della lunghezza di 50-200 bp, le versioni più recenti possono sequenziare anche 85GB in un unico ciclo di sequenza.

Un'altra piattaforma è la HiSeq dotata di due laser e quattro filtri di rilevazione dei nucleotidi, la metodica è la stessa del Genome Analyzer e può produrre fino a 600Gb di dati per corsa, la versione HiSeq 2500 può sequenziare l'intero genoma umano in un giorno.

Il sistema illumina è il più potente rispetto al Roche 454 e al sistema Solid in termini di output e il meno caro in termini di costo.

La piattaforma MISeq, rappresenta un esempio di PGM(personal genome machine), sequenziatori di piccole dimensioni adatti a piccoli laboratori e alla pratica clinica, che permettono di far fruire i vantaggi di un sequenziamento NGS in termini di velocità e costi ma con un output inferiore, circa 15Gb per corsa.

Applied Biosystems: SOLiD

Nel 2005 George Church [15] utilizzò per risequenziare il genoma di Escherichia Coli, una tecnologia basata su reazioni di ligazione, questa, opportunamente migliorata, fu poi messa in commercio nel 2007 dalla Applied Biosystsems con il nome di SOLiD (Sequencing by Oligonucleotide Ligation and Detection).

La fase di preparazione della libreria è simili a quella seguita dalla tecnologia 454, in cui i frammenti di DNA sono immobilizzati su biglie e clonati mediante PCR in emulsione; queste biglie sono poi fissate sulla superficie di vetro funzionalizzata di una cella a flusso sulla quale avviene poi il sequenziamento.

La differenza fra la piastra PicoTiter della tecnologia 454 e il supporto in vetro utilizzato in questa piattaforma è legata al fatto che l'assenza di cellette su quest'ultimo pone come unico limite al numero di sferette solo il diametro delle stesse.

Ogni ciclo di sequenziamento necessita di un primer degenere, una ligasi e quattro sonde dNTP composte da ottameri con due basi fissate e sei degeneri(le ultime tre rimovibili) e con un marcatore fluorescente terminale

Nella prima fase il primer si ibridizza con la sequenza adattatrice che lega il template alla sferetta, dopodichè la ligasi permette l'ibridazione di una sonda, con conseguente emissione di fluorescenza da parte del marcatore; all'ottamero ligato vengono poi rimosse le ultime tre basi assieme alla marcatura per le successive ripetizioni.*(Figura 4)*

A ciascuna coppia di basi è associato un colore, ma non è una marcatura univoca in quanto i colori sono 4 e le coppie di due basi possibili sono 16.

Questo sistema, anche se apparentemente sembra meno preciso del sequenziamento 1-1, in realtà permette di discriminare adeguatamente se si tratta di un errore di sequenza oppure se è presente uno SNP, oppure una delezione o inserzione.

Ogni ciclo di sequenziamento prevede che l'amplicone sia testato con 7 primer degeneri, dopodiché il frammento nuovo viene denaturato e viene ibridato sullo stesso amplicone un primer che ibridizza shiftato di una base, cosi via con altri primer shiftati fino a completare il filamento.

Con questo sistema ogni base viene interrogata due volte, riducendo all'0,01% la possibilità di errore, il limite fondamentale però è il fatto che non possono essere sequenziate “reads” superiori ai 35-40bp.[24]

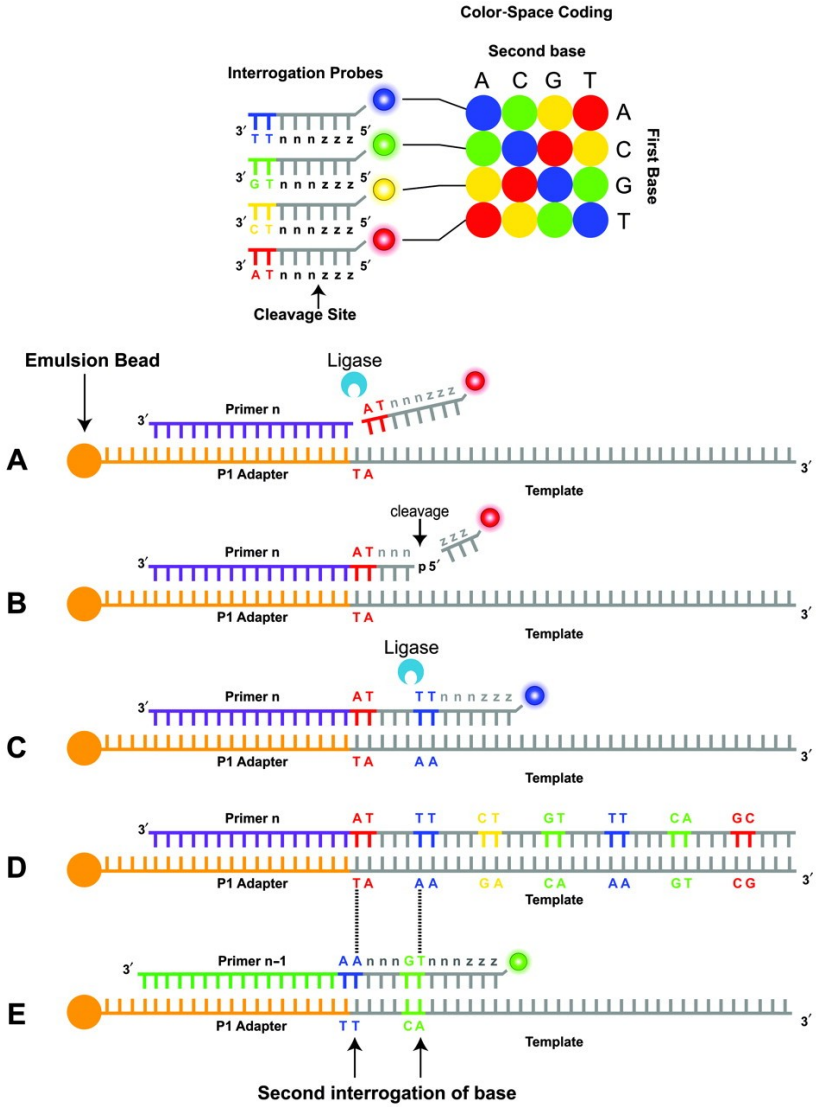

Figura 4:Sequenziamento mediante piattaforma SOLiD*(Next-Generation Sequencing: From Basic Research to Diagnostics)*

Ion Torrent

I sequenziatori Ion Torrent sono sequenziatori "benchtop" ovvero una *PGM* (*personal genome machines*), un piccolo sequenziatore con un output inferiore

alle grandi macchine adatto ai piccoli laboratori.

Furono inizialmente prodotti dalla Ion Torrent e successivamente acquistati dalla Life Technologies, attualmente sono in commercio due diverse piattaforme la Ion PGM e la Ion Proton; entrambe le due piattaforme sfruttano la tecnologia del sequenziamento tramite semiconduttore.

Praticamente il metodo si basa sulla rilevazione di ioni di idrogeno che vengono rilasciati durante la sintesi del DNA.

Il flusso di lavoro di questa piattaforma è simile alle altre viste in precedenza, la prima fase consiste nella preparazione della libreria di DNA con frammentazione del DNA template e l'attacco di appositi adapter al 5', successivamente i frammenti vengono immobilizzati su microsfere (IonSphereTM) e amplificati con PCR in emulsione.

Queste microsfere vengono depositate sui pozzetti di un chip e sequenziate, il sequenziamento consiste nel trattare con una soluzione contenente un singolo dNTP e la polimerasi il template, se avviene l'incorporazione la reazione di legame comporta la liberazione di uno ione H^+ che porta ad un'alterazione del pH che viene registrata; su queste registrazioni viene elaborata la sequenza. Se vengono incorporati due nucleotidi, vi sarà una variazione di pH "doppia".

I vantaggi principali dell'Ion Torrent è innanzitutto l'assenza di sistemi di rilevazione basati su fluorescenza che permette di ottenere una maggiore precisione, in particolare nelle zone di basi ripetute.

Inoltre con la piattaforma Ion PGM si riesce a lavorare con una quantità di DNA di partenza irrisoria circa 10ng, si riesce a sequenziare "reads" che vanno

dai 35 ai 400bp e in circa 16 ore è possibile analizzare 8 campioni in parallelo, rendendo questo sistema estremamente idoneo per l'utilizzo anche in urgenza.

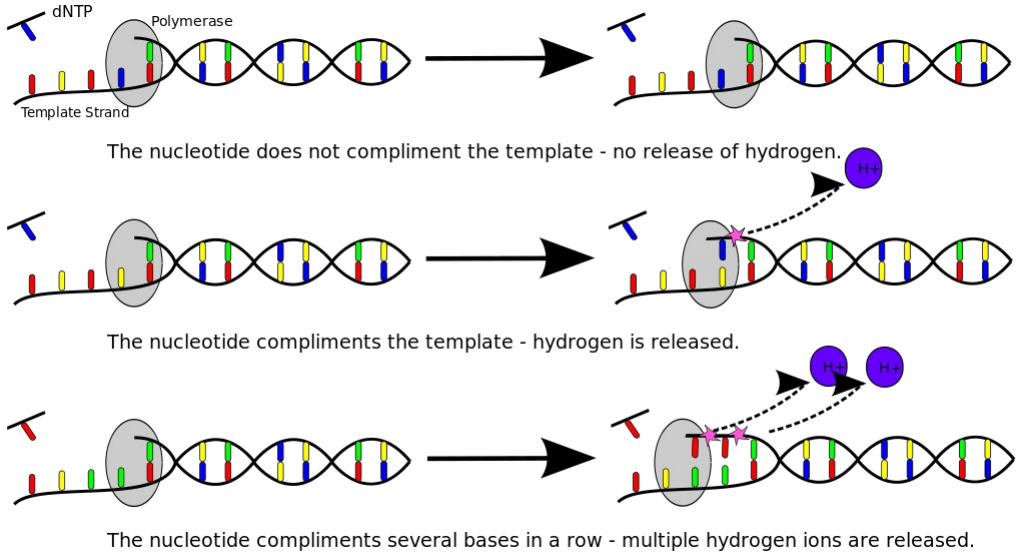

Figura 5:Principio chimico tecnologia Ion torrent*(DNTP nucleotide incorporation events David Track)*

NGS : METODICHE DI 3° GENERAZIONE

Le metodiche di 3°generazione sono anche dette metodiche a singola molecola di DNA, infatti la caratteristica principale di queste metodiche è proprio quella di non avere necessità di una amplificazione del segmento da sequenziare, producendo una serie di vantaggi nell'utilizzazione.

I vantaggi si possono riassumere in una netta riduzione dei tempi di risposta in quanto si riduce tutta la fase di preparazione del template da sequenziare, nella possibilità di elaborare "reads" più lunghe al fine di riuscire a migliorare l'accuratezza del sequenziamento *de novo*, nella maggiore efficienza nella costruzione della sequenza perché vengono eliminati gli errori che potrebbero

originarsi durante la fase di amplificazione e nella possibilità di utilizzare una piccola quantità di materiale di partenza, aspetto particolarmente utile nella pratica clinica.[19]

Le piattaforme attualmente in commercio per il sequenziamento di terza generazione sono Pacific Biosciences Single Molecule Real-Time (SMRTTM) e la piattaforma Helicos Biosciences true Single Molecule Sequencing (tSMS).

La piattaforma Helicos Biosciences true Single Molecule Sequencing (tSMS) ha molte delle caratteristiche del sequenziamento NGS ma si distingue dalle piattaforme di 2°generazione per la capacità di sequenziare direttamente l'RNA, mentre la piattaforma Pacific Biosciences Single Molecule Real-Time rappresenta un novità in quanto permette la rivelazione real-time della sequenza, ma presenta tuttavia un'accuratezza minore e un maggiore numero in percentuale di errori (>5%).

PacBio RS

Il sistema PacBioscense Single Molecule Real-Time (SMRTTM) fu per la prima volta commercializzata nel 2011 dalla Pacific Bioscences, la versione pacBio RS II è stata commercializzata nel 2013.

In questa piattaforma il sistema di sequenziamento prevede che il DNA da

sequenziare venga ancorato, insieme ad una DNA polimerasi sul fondo di una cella definita ZMW(zero-mode waveguide), che consiste in un minuscolo tubo all'interno del quale è possibile rilevare cambiamenti infinitesimali di luce colorata.[24]*(Figura 6-7)*

Questo complesso DNA-DNApolimerasi viene cimentato con soluzioni contenenti i quattro nucleotidi marcati con fluorocromi differenti, a differenza degli altri sistemi di rilevazione con fluoroforo (es.Sanger) il marcatore è posto sul fosfato e non sulla base azotata.

Man mano che la polimerasi incorpora il nucleotide, la reazione porta alla liberazione del pirofosfato che provoca un'emissione luminosa.

Il sistema di rilevazione registra l'emissione luminosa e elabora la sequenza.

L'innovazione di questo sistema è legata al fatto che non è necessaria un'amplificazione pre-sequenza riducendo al minimo eventuali errori, permettendo di sequenziare “reads” che oscillano dalle 4000-8000bp e riducendo notevolmente I tempi, un sequenziamento completo si ha in 8-10 ore.[19]

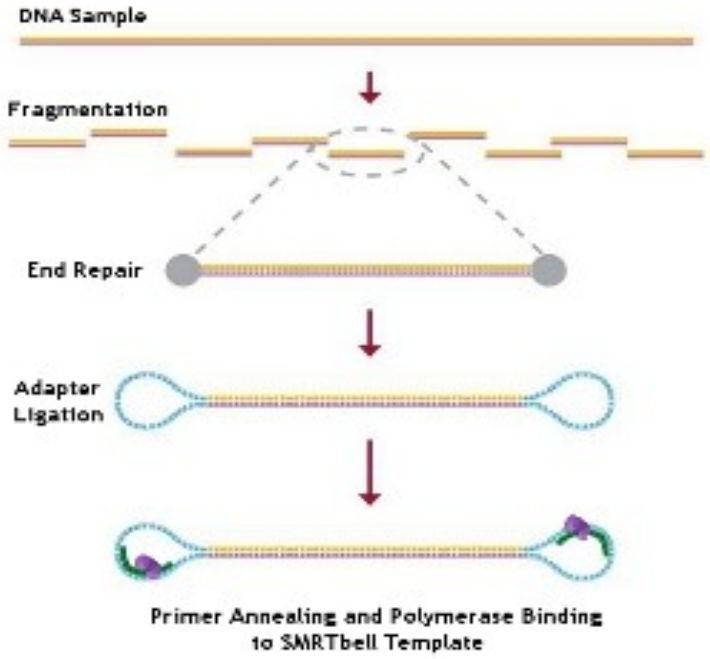

Figura 6:Preparazione template PacBios *(.umich.edu/~caparray/products/ngs/pacbio)*

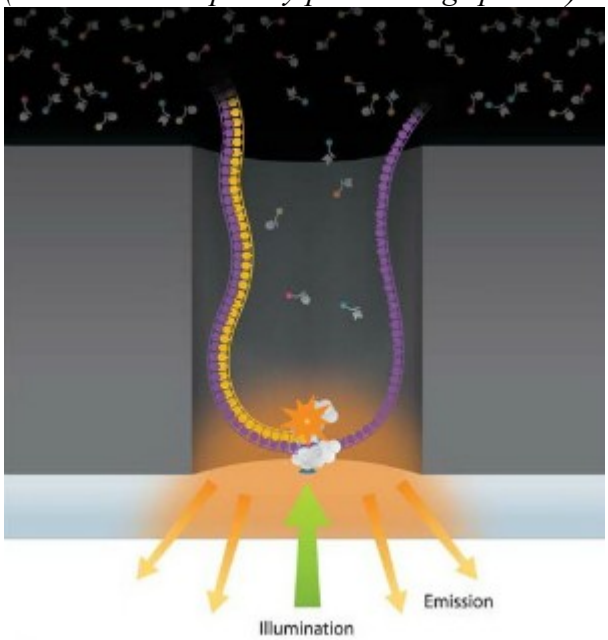

Figura 7:Cella ZMV*(PacBiosciences)*

Heliscope Sequencer

La piattaforma Heliscope Sequencer commercializzata dalla Helicos BioSciences nel 2008 è stata la prima tecnologia per il sequenziamento ad eliminare l'amplificazione clonale dei frammenti di DNA.

Il metodo si basa sulla frammentazione del DNA stampo e sulla poliadenilazione all'estremità 3', in cui l'adenosina terminale risulta marcata con un fluroforo.[16]

A questo punto i filamenti poliadenilati vengono denaturati e ibridati con code di polidT immobilizzate su celle di flusso, permettendone l'ancoraggio.

Vengono rilevate da una camera CCD le posizioni dei singoli filamenti e viene rimosso il fluoroforo, a questo punto inizia il sequenziamento mediante l'aggiunta nella cella a flusso di una soluzione contenente DNA polimerasi e uno dei 4dNTP marcati con un altro fluoroforo il Cy5; se avviene l'incorporazione viene rilevata la fluorescenza, dopodichè viene effettuato un lavaggio e il filmaneto di DNA testato con una soluzione di un altro nucleotide.

Ogni ciclo di test con i 4 nucleotidi viene definito “quad” solitamente una seduta di sequenziamento prevede lo svolgimento di 25-30 quad per ottenere reads di 40-50bp.(Figura 8)

I principali vantaggi di questa tecnica sono legati all'assenza di amplificazione clonale che riduce i tempi di indagine e riduce la possibilità di incorporare errori, ma è stato dimostrato una scarsa accuratezza nei tratti di basi ripetute, per questo motivo sono stati brevettati dalla Helicos dei terminatori virtuali che

riducono la processività della DNA polimerasi in modo che vengono aggiunte singole basi.

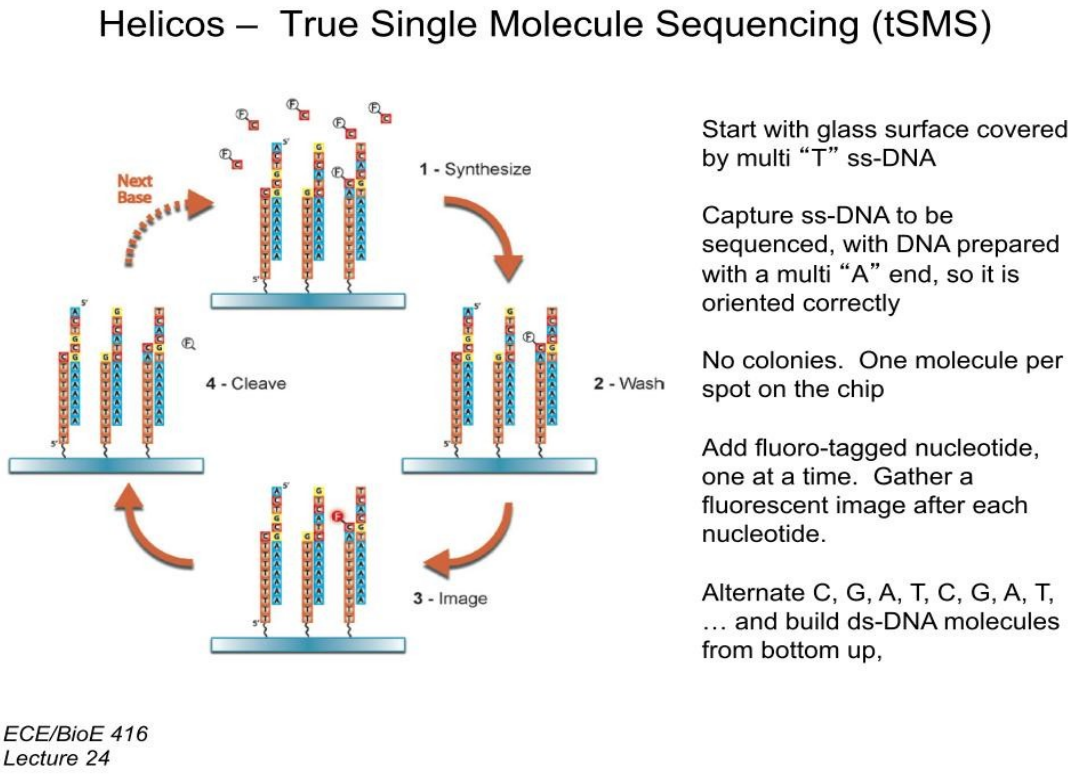

Figura 8 : Helicos workflow*(Brian Cunningham NanoBio Node)*

NGS:METODICHE DI 4°GENERAZIONE

Tecnologia a nanopori

Nell'ultimo decennio i ricercatori hanno messo a punto un sistema di sequenziamento basato sull'utilizzo di tecnologia a nanopori.[19]

In questo sistema una membrana composta da nanopori è posta tra due camere riempite di una soluzione elettrolitica, applicando una corrente elettrica la molecola di DNA viene fatta passare all'interno di questa apertura, il passaggio

della molecola modifica il flusso di corrente ionica attraverso il poro.

L'osservazione fondamentale, esposta per la prima volta da Church, Deamer, Barton, Baldarelli e Kasianowicz nel 1995, è che ogni base crea una differente alterazione nella corrente ionica attraverso il nanoporo quindi, potendo valutare queste alterazioni è possibile riconoscere ogni singola base.

La tecnologia a nanopori presenta notevoli vantaggi in quanto non richiede marcature dei "reads", non richiede amplificazione dei frammenti, ha un approccio di sequenziamento a singola molecola con elevato throughput, richiede piccole quantita di materiale di partenza; i principali problemi che si sono dovut affrontare sono stati la velocità di passaggio della molecola e la larghezza dei nanopori.

Nel tentativo di risolvere questi problemi e di ottimizzare i risultati sono stati progettati vari tipi di nanopori, biologici, struttura su supporto solido e strutture ibride tra i due.

Le strutture a nanopori biologiche principalmente sperimentate sono l'a-emolisina e la porina A (MspA), l'utilizzo di queste strutture è dovuto principalmente al fatto che si possono creare delle strutture di diametro ridotto circa 3,6nm e perchè gli ultimi studi dimostrano come l' a-emolisina sia perfettamente adatta a distinguere i frammenti di DNAss dai frammenti a RNA rendendolo uno strumento efficiente per il sequenziamento.

Il limite principale dell'utilizzo di nanopori biologici è la velocità con cui il DNA passa attraverso, infatti gli studi hanno rilevato una velocità di 1 nucleotide per microsecondo;questa velocità fa si che solo un piccolo numero

di ioni (circa 100) sono presenti nel nanoporo per rilevare le variazioni di correnti del passaggio del DNA.

Ma queste piccole variazioni di corrente che si vengono a generare vengono coperte dalle normali variazioni termodinamiche impedendo la valutazione e la rilevazione del passaggio della base.

Gli approcci per risolvere questa problematica sono stati di tipo attivo o passivo, dal punto di vista dell'approccio attivo al problema, un modo di rallentare il flusso di basi di DNA attraverso il nanopore è stato l'utilizzo di un complesso enzimatico che pur permettendo il movimento elettroforetico attraverso il nanoporo è in grado di rallentare e controllare il movimento della molecola.

L'approccio passivo, più semplicemente richiede, la marcatura del filamento con un uncino che sfrutta la carica del nanoporo per ancorarsi sulla sua superficie; sono tuttavia metodiche che devono essere ancora indagate e perfezionate.[20]

ANALISI DEI DATI

I test eseguiti con piattaforme NGS generano una quantità di informazioni senza precedenti, infatti, a fronte di una riduzione dei tempi e dei costi di lavoro, risultano essere notevolmente complesse le procedure di acquisizione e di elaborazione dei dati, soprattutto per la mole di informazioni che sono in grado di generare.

Questa massa di informazioni richiede un sistema chiamato “data-pipeline system”che ha la capacità di archiviare,gestire ed elaborare l'enorme mole di dati; per avere un'idea una singola seduta analitica di una piattaforma 454 GS LX produce 15GB di dati, Illumina 1TB e SOLiD 15TB.[25]

Una delle operazioni principali nell'elaborazione dei dati è chiamata “base calling” e consiste nel convertire le immagini acquisite in “reads”

Mediante l'impiego di algoritmi specifici per ogni piattaforma vengono valutati alcuni parametri della sequenza, quali intensità, rumore di fondo e presenza di eventuali segnali aspecifici; tutto ciò è finalizzato ad assegnare dei punteggi definiti “quality scores” correlati alla probabilità di errore.

L'esistenza dei “quality scores” è fondamentale in quanto permettono di escludere “reads” o eliminare basi che presentano una bassa qualità, per migliorare l''accuratezza dell'allineamento e per determinare sia la sequenza consensus sia le eventuali varianti.

Vi è una maggiore difficoltà nell'allineamento e l'assemblaggio delle “reads” su piattaforme NGS rispetto a dati ottenuti con metodo Sanger, ciò è dovuto alla minore lunghezza delle prime.[26]

L'allineamento e l'assembly presenta un limite importante che è quello di non permettere un allineamento delle sequenze univoco in presenza di reads molto brevi oppure quando si devono risequenziare genomi molto lunghi e complessi.

In numero possiamo dire che il genoma di Escherichia Coli può essere allineato al 97% in modo univoco utilizzando “reads” da 18bp mentre un genoma umano solo al 90% con reads da 30bp.

L'allineamento univoco è ridotto dalla presenza di sequenze ripetute e dall'omologie di sequenze condivise e pseudogeni, i software risolvo questi problemi di allineamento lasciando alcuni gaps oppure assegnano le reads a posizioni multiple.[27]

Nel sequenziamento *de novo* le reads non univoche saranno escluse; questo giustifica la necessità di scegliere una piattaforma adatta per il tipo di sequenziamento da svolgere.

L'accuratezza dell'analisi per le tecnologie NGS è migliorata attraverso il sequenziamento ripetuto di una data regione di interesse, cosi da raggiungere un elevata copertura della sequenza ("coverage") in materia tale da costruire una sequenza consensus.

Il "coverage" deve essere molto elevato cosi da permettere di rilevare variazioni nucleotidiche sulla sequenza.

Alcuni studi su genoma di lievito hanno messo in luce come la possibilità di rilevare variazioni nucleotidiche si può ottenere quando i valori di coverage sono superiori a 15.

Attualmente sono stati messi in commercio numerosi software per l'allineamento e l'assembly che utilizzano come sistema operativo Linux o Windows, ogni software è specificatamente indicato per un tipo di lavoro.[28]

Una situazione particolare è legata ai software di interpretazione per gli approcci RNA-seq e ChIP-seq (*vedere paragrafi successivi*).

Il software per la tecnologia RNA-seq ha rappresentato una sfida importante in quanto era necessario ottenere sia l'allineamento dei trascritti in cui è avvenuto

lo splicing sia a livello delle code di poli(A).

Il software attualmente in uso permette di identificare sequenze "consensus" per i siti di splicing e delle regioni di giunzione introne-esone con un "basso" coverage di allineamento.[29]

Per quanto riguarda la metodica ChIP-Seq i dati ottenuti presentano dei picchi caratteristici dei siti di legame delle proteine, l'identificazione di questi picchi permette di mappare il sito di legame della proteina.[30]

AMBITI DI APPLICAZIONE DELLE TECNICHE NGS

L'avvento delle tecnologie NGS ha notevolmente accelerato la crescita di vari settori di ricerca genomica, consentendo di effettuare esperimenti che in precedenza presentavano notevoli ostacoli soprattutto da un punto di vista

economico.

ANALISI GENOMICA

L'analisi di un intero genoma, da quelli di microrganismi a quelli umani[28,31-32], è un procedimento che richiede un'elevata processività per l'elevata quantità di materiale genetico da decifrare; i dati relativi alle piattaforme di seconda generazione indicano che queste siano in grado di sequenziare un intero genoma in circa 10 giorni a fronte degli anni che sono stati necessari utilizzando la tecnica di Sanger.

È stato possibile sequenziare il genoma di cellule citogeneticamente normali derivanti da un paziente con leucemia mieloide acuta per identificare nuove mutazioni genetiche tumore specifiche.[33]

In particolare la tecnologia 454, permettendo il sequenziamento di reads più lunghe rispetto alle piattaforme SOLiD e Illumina, risulta particolarmente indicata per il sequenziamento *de novo*, cioè senza una sequenza genomica di riferimento.[28]

Inoltre la possibilità di analizzare reads più lunghe permette inoltre ottenere informazioni all'interno di zone lunghe alcune centinaia di basi che presentano sequenze ripetute e potrebbero creare problemi di allineamento.

RISEQUENZIAMENTO

Alcuni particolari tipi di studio richiedono il sequenziamento di sottoregioni

genomiche o di gruppi di geni.

Questo può essere necessario per identificare mutazioni e polimorfismi associati a malattie, correlate all'insorgenza di neoplasie, oppure per l'identificazione di regioni che hanno un coinvolgimento in malattie genetiche mediante studi di linkage disequilibrium, oppure per la definizione di particolari aplotipi.[34,35]

Un uso più efficace delle tecnologie NGS in questo settore non può prescindere da step di arricchimento genomico, volto a creare un singolo target di sequenza.

Un sistema, utilizzato quando la zona da sequenziare è intorno alle centinaia di Kb, è la "long-range PCR"(utilizzata per esempio nella tipizzazione HLA), mentre per regioni più grandi solitamente l'approccio consiste in una frammentazione del DNA genomico con conseguente cattura ad opera di oligonucleotidi sonda complementari a regioni di interesse.

ANALISI DEL TRASCRITTOMA

Le tecnologie NGS hanno fornito un nuovo e potente sistema, denominato "RNA-Seq" per la mappatura e la quantificazione dei trascritti nei campioni biologici.

Lo studio del trascrittoma permette di determinare le differenze di espressione di un gene nei vari tessuti, o nello stesso tessuto, in diverse fasi di sviluppo

cellulare e dell'organismo.

Si possono poi studiare le variazioni di espressione genica in differenti condizioni ambientali, in risposta a meccanisimi di regolazione genetica e epigenetica.

Da un punto di vista operativo l'RNA totale è isolato e convertito in cDNA attraverso una retrotrascrizione mediata da primer e la generazione successiva del secondo filamento di cDNA ad opera di una RnasiH e di una DNA polimerasi.

Il DNA viene poi solitamente sottosto ad una pcr di arricchimento e poi frammentato, legato ad adattatori e analizzato con tecnologie NGS.

Le "reads" così generate vengono allineate ad un genoma di riferimento, comparate a sequenze di trascritti noti oppure utilizzate per un "de novo assembly" per costruire una mappa di trascrizione nuova.

L'approccio RNA-seq permette dei notevoli vantaggi rispetto all'utilizzo degli array nell'analisi della trascrizione genica, innanzitutto con gli array è possibile identificare solo sequenze genomiche note, inoltre l'approccio RNA-seq permette di ottenere una risoluzione a livello della singola base e permette di distinguere diverse isoforme di RNA, di determinare l'espressione allelica e di rilevare variazioni di sequenza.[36]

I livelli di espressione sono dedotti dal numero di reads totale normalizzato per la lunghezza degli esoni che possono essere mappati in modo univoco.

Il sistema RNA-seq viene utilizzato per confermare e aggiornare le informazioni geniche nelle re

gioni di giunzione tra esone e introni, coinvolte nei fenomeni di splicing alternativo, mappando le reads nei siti giunzione GT-AG è possibile ottenere sia informazioni di tipo quantitativo che qualitativo.[36]

MAPPATURA DELLE PROTEINE LEGANTI DNA E ANALISI CROMATINA

La tecnologia ChIP-on chip unisce l'immunoprecipitazione della cromatina (ChIP) con la tecnologia dei microarray(on chip); questa tecnica è stata particolarmente utilizzata per identificare i siti di legame dei fattori di trascrizione, delle proteine istoniche o delle regioni di replicazione.[37]

In questo tipo di analisi le proteine associate al DNA sono ibridate chimicamente al loro sito di legame e il DNA è frammentato, successivamente le proteine subiscono un' immunoprecipitazione e il DNA viene ibridato a un "array"di sequenze oligonucleotidiche.

Sono molti gli approcci sperimentali che sostituiscono la tecnica ChIP-on chip con la tecnica ChIP-Seq in cui il DNA che viene recuperato a seguito dell'immunoprecipitazione viene analizzato con piattaforme NGS.

Le "reads" ottenute sono mappate al genoma di interesse per costruire una mappa dei siti di legame delle proteine al DNA.[38]

Analogamente all'approccio RNA-seq, presenta l'enorme vantaggio di poter individuare anche nuovi siti di legame DNA-proteine, perché non necessita di un confronto con una sequenza conosciuta.

METAGENOMICA

La metagenomica è una scienza che segue come approccio lo studio di comunità microbiche direttamente nel loro habitat naturale, eliminando la difficoltà della raccolta e della coltivazione in laboratorio, in quanto spesso si tratta di organismi che vivono in condizioni ambientali estreme o di patogeni.

Si basa sul sequenziamento di un insieme di microrganismi permettendo di elaborare una quantità di materiale genetico definito metagenoma.

Da un punto di vista operativo, viene prelevato un campione e il DNA presente viene sequenziato, il risultato è la presenza di DNA provenienti da organismi differenti che vivono nello stesso habitat e quindi permettere

di valutare analogie e somiglianze; soprattutto di identificare particolari geni necessari alla sopravvivenza.

Esempi di studi di metagenomica sono le analisi di popolazioni nei fondi oceanici, la caratterizzazione della microflora della cavità orale umana.[39,40]

L'impatto delle tecnologie NGS sulla metagenomica è stato enorme, infatti il DNA prelevato da un campione viene frammentato e sequenziato e le sequenze ottenute possono essere allineate con sequenze di riferimento di microrganismi che si ipotizza presenti sul campione.

Inoltre è possibile fare una valutazione sulle analogie di sequenza, definendo specie strettamente correlate o specie filogeneticamente distanti; l'analisi "de novo assembly" permette di porre le basi per l'individuazione di specie

potenzialmente nuove.

LA maggior parte degli studi di metagenomica viene fatta con piattaforme NGS con tecnologia 454 in quanto la possibilità di ottere reads più lunghe facilità l'allineamento dei genomi microbici e la costruzione di sequenze de novo in caso di genomi microbici non ancora caratterizzati.

FARMACOGENETICA e FARMACOGENOMICA

La conoscenza più approfondita del Genoma Umano e la possibilità di sequenziare anche regioni molto estese del DNA ha dato un forte impulso allo sviluppo di alcune branche della genetica e della genomica.

La farmacogenomica è una disciplina che si pone come obiettivo lo studio dei geni che modulano la risposta farmacologica ed è principalmente orientata alla ricerca di nuovi bersagli terapeutici, allo sviluppo di farmaci e allo studio della risposta ai farmaci.

La farmacogenetica, come sottoinsieme della farmacogenomica, studia come le variazioni sul DNA influenzano la risposta al farmaco, le reazioni al dosaggio e le interazioni farmaco-farmaco.

Sono stati documentati diversi casi di reazioni avverse, anche letali, a farmaci in alcuni individui legati a singoli polimorfismi genetici negli enzimi deputati al metabolismo dei farmaci; si tratta principalmente di SNPs (Single Nucleotide Polimorphism), variazioni del numero di copie (CNV) o di ripetizioni in tandem (VNTR) o di microsatelliti.

L'avvento delle tecnologie NGS ha permesso di portare avanti progetti come HapMap e 1000 genome creando un database di genomi al fine di fornire supporto per l'identificazione delle varianti alleliche legate a malattie o ad una diversa risposta ai farmaci ed a caratterizzare polimorfismi genetici a bassa frequenza(<1%).[41]

CONCLUSIONI

Con il completamento nel 2003 del Progetto Genoma Umano sono stati messi i in luce alcuni punti cruciali relativi alla codifica del genoma umano:

- Il numero esiguo di geni, contrariamente alle previsioni, che non possono da soli spiegare le differenze tra gli organismi
- Il 97% del materiale genetico era ancora a funzione sconosciuta
- I tempi troppo lunghi, necessari alla codifica di un intero genoma umano
- I costi troppo elevati

Conseguentemente a ciò mentre da un lato gli approcci sperimentali si muovevano nella direzione di valutare le interazione tra geni, di andare a cercare mutazioni del DNA che permettessero di individuare le origine della malattie, dall'altro i tempi e i costi rallentavano fortemente questo tipo di studi.

Le ricerche scientifiche si sono quindi indirizzate verso la messa a punto di piattaforme che potessero ridurre i costi e i tempi per il sequenziamento e parallelamente potessero produrre risultati di qualità elevata.

Negli ultimi anni la tecnologia relativa al sequenziamento ha compiuto notevoli passi avanti, sia per il Sanger Sequencing in cui i sequenziatori a 16 e a 24 capillari hanno ridotto notevolmente i tempi di lavoro ma soprattutto con l'avvento delle tecnologie NGS con un nuovo modo di concepire il sequenziamento in maniera massiva e parallela.

Le tecnologie successive più recenti a singola molecola e le tecnologie ancora in fase di sviluppo, quali l'utilizzo di nanopori per il sequenziamento dimostrano come il processo sia ancora in piena evoluzione.

Inoltre anche le tecnologie già presenti beneficiano del fatto di essere sistemi ancora in una fase di evoluzione che riguarda i sistemi di automazione, il perfezionamento della chimica dei reagenti, la riduzione dei costi e il miglioramento della gestione dei dati.

Nel 2008 il sequenziamento dell'intero genoma umano con la tecnologia Roche 454 ha richiesto solo 5 mesi di lavoro con un costo di circa 1,5 milioni di dollari.[42]

Attualmente il costo per le analisi effettuate mediante piattaforme NGS

richiede un investimento economico importante sia per quanto riguarda le attrezzature, sia per i reagenti, ma rimane sostanzialmente inferiore a quello richiesto dal sequenziamento di Sanger(vedi tabella) mentre la mole di dati prodotta è nettamente maggiore con minor tempo impiegato.

Altri limiti a cui ancora la tecnologia NGS deve far fronte è il flusso di lavoro per la preparazione delle librerie di DNA e l'analisi dei dati che risulta essere ancora complessa e richiede elevate conoscenze di bioinformatica.

L'introduzione di tecnologi single-molecule si muove nella direzione di ridurre i tempi necessari per la processazione.

Attualmente l'impatto maggiore delle tecnologie NGS è stato sulla ricerca di base, ma si sta lentamente passando all'utilizzo di questo approccio anche nella diagnostica clinica.

L'immissione nel mercato dei PGM, ovvero delle persone genome machine, si sta muovendo proprio in questa direzione.

SI tratta di apparecchiature più piccole, specificatamente indicate per il sequenziamento di sequenze specifiche (target-sequencing) particolarmente indicate per l'utilizzo nella pratica clinica.

Queste piattaforme hanno costi inferiori rispetto alle attrezzature NGS in grado di sequenziare l'intero genoma, ed hanno anche un throghput inferiore ma sono in grado di fornire dati di sequenza per la pratica clinica quali diagnosi molecolare di malattia, tipizzazione tissutale, screening per la predisposizione genetica.

Ci vorrà ancora tempo perché avvenga il passaggio alla diagnostica clinica, ma

le prospettive sono ottimistiche sia in termini di costi, di tempi e di applicazioni possibili.[43].

BIBLIOGRAFIA

1. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M, ***What is a gene, post-ENCODE? History and updated definition in Genome Research,*** *vol. 17, nº 6, 2007, pp. 669–681, DOI:10.1101/gr.6339607, PMID 17567988.*
2. International Human Genome Sequencing Consortium, ***Finishing the euchromatic sequence of the human genome.*** *in Nature, vol. 431, nº 7011, 2004, pp. 931–45.*
3. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter

JC. ***The diploid genome sequence of an individual human.*** *Biol. 2007 Sep 4;5(10):e254.*

4. Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, Rosenbloom KR, Zweig AS, Haussler D, Kent WJ. ***Comparative genomic analysis using the UCSC genome browser.*** *Methods Mol Biol. 2007;395:17-34.*
5. Venter JC & other ***The sequence of the human genome****. Science. 2001 Feb 16;291(5507):1304-51.*
6. Gilbert W, Maxam AM***. The nucleotide sequence of the lac operator*** *Proc Natl Acad Sci U S A. 1973 Dec;70(12):3581-4.*
7. Maxam AM, *Gilbert* ***A new method for sequencing DNA****. W.Proc Natl Acad Sci USA 1977;74:560-4*
8. Sanger F.,Nicklen S.,Coulson AR***. DNA sequencing with chan-terminating inhibitors.*** *Proc Natl Acad Sci USA 1977;74:5463-7*
9. Strasser BJ. ***The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine.*** *Isis. 2011 Mar;102(1):60-96.*
10. Kary B. Mullis - ***Nobel Lecture: The Polymerase Chain Reaction".*** *Nobelprize.org. Nobel Media AB 2014. Web. 3 Sep 2014.*
11. *http://tools.lifetechnologies.com/content/sfs/posters/ABI6247_SOLiD_Timeline_v4_ONLINE.pdf*
12. Sequencing Consortium. ***Genome sequence of the nematode C. elegans: a platform for investigating biology*** *Science. 1998 Dec 11;282(5396):2012-8.*
13. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E. ***The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing.*** *PLoS One. 2007 Feb 14;2(2):e197.*
14. Bentley DR. Balasubramanian S,Swerdlow HP. ***Accurate whole human***

genome sequencing using reversible terminator chemistry. *Nature 2008;456:53-9*

15. Shendure J, Porreca GJ, Reppas NB. ***Accurate Multiplex polony sequencing of an evolved bacterial genome.*** *Science 2005;309:1728-32*
16. Braslavsky I, Herbert B, Kartalov E. ***Sequence information can be obtained fronm single DNA molecules***. *Proc Natl Acad Sci USA 2003;100:3960-4*
17. Xu M, Fujita D, Hanagata N. ***Perspectives and challenges of emerging single-molecule DNA sequencing technologies***. *Small. 2009 Dec;5(23):2638-49. doi: 10.1002/smll.200900976.*
18. Venkatesan BM, Bashir R. ***Nanopore sensors for nucleic acid analysis.*** *Nat.Nanotechonol.6,615-624(2011)*
19. Che-Seng Ku, Roukos DH. ***From next-generation sequencing to nanopore sequencing technology:paving the way to personalized genomic medicine.*** *Expert Rev.Med.Devices 10(1),1-6(2013)*
20. Nyren P, Pettersson B***. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay.*** *Anal Biochem 1993;208:171-5*
21. Tawfik DS,Griffiths ***AD Man-made cell-like compartments for molecular evolution***. *Nat.Biotechnol 1998;16:652-6*
22. Pearson BM, Gaskin DJ, Segers RP, et ***al The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828).****J Bacteriol 2007;189:8402-3*
23. Huse SM, Huber JA, Morrison HG ***Accuracy and quality of massively parallel DNA pyrosequencing.*** *Genome Biol 2007;8:R143*
24. Karl V .Voelkerding, S.A.Dames, J.D.Durtschi. ***Next Generation Sequencing: from basic research to diagnostics.*** *Clin Chem 2009;55:641-58*
25. Pop M, Salzberg SL.Trends ***Bioinformatics challenges of new***

sequencing techonology. *Genet 2008;24:142-9*

26. Rougemont J,Amzallag A, Iseli C ***Probabilistic base-calling of Solexa sequencing data.*** *BMC Bioinformatics 2008;9:431*
27. Whiteford N, Hasham N, Weber G. ***An analysis of the feasibility of short read sequencing.*** *Nucleic Acid Res 2005;33:e171*
28. Smith DR, Quinlan AR, Peckham HE. ***Rapid whole-genome mutational profiling using next-generation sequencing technologies.*** *Genome Res 2008;1818:1638-42*
29. Wang Z, Gerstein M, Snyder M ***RNA-seq:a revolutionary tool for trasciptomics..*** *Nat Rev Genet 2009;10:57-63*
30. Valouev A.Johnson DS, Sundquist A ***Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.*** *Nat Methods 2008;5:829-34*
31. Satkoski JA, Mahli R, Kanthaswamy S. ***Pyrosequencing as a method for SNP identification in the rhesus macaque (Macaca Mulatta)*** *BMC Genomics 2008;9:256*
32. Wang J, Wang W, Li R ***The diploid genome sequence of an Asian individual..*** *Nature 2008;456:60-5*
33. Ley TJ, Mardis ER, Ding L***. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.*** *Nature 2008;456:66-72*
34. Yeager M, Xiao N, Hayes RB. ***Comprehensive resequence analysis of a 136kb region of human chromosome 8q24 associated with prostate colon cancers.*** *Hum Genet 2008:124:161-70*
35. Ding l, Getz G,Wheeler DA. ***Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008;455:1069-75***
36. Wang Z, Gerstein M,Snyder M. ***RNA-seq: a revolutionary tool for transcriptomics.*** *Nat.Rev.Genet 2009;10:57-63*
37. Ren B. Robert F, Wyrieck JJ ***Genome-wide location and function of DNA binding proteins.*** *Science 2000;290:2306-9*
38. Barski A, Cuppadah S, Cui K. ***High-resolution profiling of histone***

methylations in the human genome. *Cell 2007;129:823-37*

39. Huber JA, Mark Welch DB, Morrison HG. ***Microbial population structures in the deep marine biosphere.*** *Science 2007;318:97-100*
40. Keijser BJ, Zaura E, Huse ***SM Pyrosequencing analysis of the oral microflora of healthy adults..*** *J Dent Res 2008;87:1016-20*
41. Russo R Capasso M, Iolascon A ***Farmacogenetica e farmacogenomica in pediatria: stato dell'arte e prospettive future.*** *Frontiere2013; 43:43-50*
42. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM .***The complete genome of an individual by massively parallel DNA sequencing.*** *Nature. 2008 Apr 17;452(7189):872-6*
43. Xuan J1, Yu Y, Qing T, Guo L, Shi L***. Next-generation sequencing in the clinic: promises and challenges****. Cancer Lett. 2013 Nov 1;340(2):284-95.*

RINGRAZIAMENTI

Alla fine di questo lungo percorso vorrei ringraziare tutte le persone che mi hanno permesso e aiutato a raggiungere questo obiettivo.

Ringrazio l'Università di Pisa, tutti i professori che ho incontrato in questi anni, in particolare il Prof.Aldo Paolicchi che anche in questa occasione mi ha seguito come relatore.

Ringrazio i miei colleghi di corso con cui abbiamo condiviso questi cinque anni di scuola.

Ringrazio anche la ditta Lagitre che, nonostante il lavoro, mi ha permesso di portare a termine la scuola.

E in ultimo vorrei ringraziare le persone che mi hanno sostenuto e supportato, ma soprattutto aiutato, in questi anni, mia madre, mio padre, mia sorella ,il mio fidanzato Piero e tutti gli altri che mi sono stati vicini.

Senza di voi non sarei mai riuscita ad arrivare fino a questo punto.

Grazie per tutto quello che avete fatto per me.